

This is a repository copy of *Neural predictors of evaluative attitudes towards celebrities*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/105595/>

Version: Accepted Version

Article:

Izuma, Keise orcid.org/0000-0003-0256-3571, Shibata, Kazuhisa, Matsumoto, Kenji et al. (1 more author) (2017) Neural predictors of evaluative attitudes towards celebrities. *Social Cognitive and Affective Neuroscience*. pp. 382-390. ISSN 1749-5024

<https://doi.org/10.1093/scan/nsw135>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Neural predictors of evaluative attitudes towards celebrities

Abbreviated Title: Decoding social attitude

Keise Izuma^{1,2,3*}, Kazuhisa Shibata^{4,5}, Kenji Matsumoto³,
and Ralph Adolphs²

¹ Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.

² Division of Humanities and Social Sciences, California Institute of Technology,
1200 E. California Blvd, Pasadena, CA, 91125, USA.

³ Brain Science Institute, Tamagawa University, 6-1-1, Tamagawa-gakuen, Machida,
Tokyo 194-8610, Japan.

⁴ Department of Cognitive, Linguistic, and Psychological Sciences, Brown University,
190 Thayer Street, Providence, RI 02912

⁵ Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-
ku, Nagoya 464-8601, Japan

* Correspondence should be addressed to Keise Izuma, Department of Psychology,
University of York, Heslington, York, YO10 5DD, UK. Email:
keise.izuma@york.ac.uk

The total number of words in the main text: 5359 words

Figure Count: 2

Table Count: 1

ABSTRACT

Our attitudes towards others influence a wide range of everyday behaviors, and have been the most extensively studied concept in the history of social psychology. Yet they remain difficult to measure reliably and objectively, since both explicit and implicit measures are typically confounded by other psychological processes. We here address the feasibility of decoding incidental attitudes based on brain activations. Participants were presented with pictures of members of a Japanese idol group inside an fMRI scanner while performing an unrelated detection task, and subsequently (outside the scanner) performed an incentive-compatible choice task that revealed their attitude toward each celebrity. We used a real-world election scheme that exists for this idol group, which confirmed both strongly negative and strongly positive attitudes towards specific individuals. Whole-brain multivariate analyses (searchlight-based support vector regression) showed that activation patterns in the anterior striatum predicted each participant's revealed attitudes (choice behavior) using leave-one-out (as well as 4-fold) cross-validation across participants. By contrast, attitude extremity (unsigned magnitude) could be decoded from a distinct region in the posterior striatum. The findings demonstrate dissociable striatal representations of valenced attitude and attitude extremity, and constitute a first step toward an objective and process-pure neural measure of attitudes.

Key Words: attitude, attitude extremity, preference, fMRI, MVPA, striatum

Attitudes comprise our evaluations of an object, a place, an idea, another person, or oneself: which are good or bad, which do we want to approach or avoid? Attitudes influence a wide range of human behaviors ranging from everyday social interactions with other people, health behavior, and political behavior to international relations. Although attitude is one of the most extensively studied concepts in the entire history of social psychology (Petty et al., 2009), it has remained a challenge how best to objectively and accurately measure people's attitudes. Problems with self-report measures have been well-documented in the past (in particular, social desirability bias) (DeMaio, 1984; Podsakoff et al., 2003). While implicit measures of attitudes (Wittenbrink and Schwarz, 2007), such as the implicit association test (IAT; Greenwald et al., 1998), have offered partial solutions to this problem, their validity is also debated (e.g., Bosson et al., 2000; Karpinski and Hilton, 2001; Fazio and Olson, 2003; Arkes and Tetlock, 2004; Conrey et al., 2005; Blanton et al., 2006; Fiedler et al., 2006; Sherman, 2009). For example, scores on the IAT are influenced by multiple cognitive processes, not only implicit attitude (e.g., Conrey et al., 2005; Sherman, 2009).

In the present study, we aimed to test the feasibility of measuring people's attitudes toward other familiar people based on the power of their brain activations to predict choice behavior, and without requiring any explicit or implicit task in the scanner. A neural measure of incidental attitudes would have significant potential to provide a process-pure metric, and avoid the contamination with many other processes that limits currently available explicit as well as implicit measures of attitude (Conrey et al., 2005; Sherman, 2009). Our approach used multi-voxel pattern analysis (MVPA) together with support vector regression on functional magnetic resonance imaging (fMRI) signals. While conventional univariate fMRI analysis

1 compares the strength of activations in each single voxel independently, MVPA
2 classifies the distributed patterns of activations across multiple voxels in a high-
3 dimensional space, and can be more sensitive for detecting and distinguishing
4 different psychological states (e.g., Vickery et al., 2011; Jimura and Poldrack, 2012).

5 In the present study, we particularly focus on people's attitudes toward social
6 objects, familiar people. Although many past social psychological studies as well as
7 social neuroscience studies on attitudes have focused on racial attitudes (e.g., Phelps
8 et al., 2000; Stanley et al., 2011; for reviews, Ito and Bartholow, 2009; Kubota et al.,
9 2012; Amodio, 2014), studies of familiar famous people offer some advantages:
10 unlike racial attitude, the relationship between self-reported attitude and behavior can
11 be more straightforward (e.g., less susceptible to social desirability bias) and thus the
12 relationship among neurally measured attitude, self-report attitude, and behavior
13 (revealed attitude) is more easily interpretable.

14 Furthermore, in addition to testing the feasibility of inferring people's attitudes
15 based on brain activations (a question of psychological interest regardless of the
16 neuroanatomical details), the present study also investigates the regional neural
17 representations of attitudes and attitude extremity (i.e., how extreme attitude is
18 regardless of its valence). While neuroeconomics studies have extensively
19 investigated neural representations of attitude toward non-social objects (e.g., foods,
20 DVDs, etc.) (for review, Levy and Glimcher, 2012; Clithero and Rangel, 2014), only
21 few studies (e.g., Cunningham et al., 2003; Knutson et al., 2006; Cunningham et al.,
22 2008; Tusche et al., 2013) investigated the neural representation of attitude toward
23 other familiar people, despite the high relevance of this topic to our everyday social
24 behaviors. Furthermore, attitude extremity is known to be one of the important
25 attitude properties (Petty & Krosnick, 1995). For example, attitude extremity affects

1 an individual's information processing (e.g., Powell and Fazio, 1984; Van Boven et
2 al., 2012) and modulates the relationship between explicit and implicit attitudes
3 (Karpinski et al., 2005). However, its neural mechanisms also remain largely
4 unexplored (see Cunningham et al., 2008; Luttrell et al., 2016, for a notable
5 exception).

6 In the fMRI scanner, participants were presented with 10 members of a Japanese
7 female idol group (the Japanese music performance group “AKB48”). After the
8 scanning session, they performed an incentive-compatible choice task, which
9 behaviorally quantified each participant's attitude toward each of the 10 members
10 (with a real-world outcome). MVPA was applied to these data to identify those brain
11 regions that could predict a participant's choice behavior toward each member. Since
12 different idol group members were in fact associated with idiosyncratic preferences
13 amongst our participants, the neural data should uniquely encode the attitude for a
14 member that predict the preference choice, unconfounded by the perceptual
15 appearance of the member. We expected to find neural signatures of social attitude
16 within those brain regions previously associated with value representations in general,
17 including striatum, ventromedial prefrontal cortex (vmPFC), insula, amygdala, and
18 anterior cingulate cortex (ACC) (e.g., Buchel et al., 1998; LaBar et al., 1998; Delgado
19 et al., 2000; Knutson et al., 2000; O'Doherty et al., 2001; Knight et al., 2005; Hare et
20 al., 2008; Izuma et al., 2008; for meta-analyses, Etkin et al., 2011; Bartra et al., 2013;
21 Sescousse et al., 2013). Nonetheless, in order to obtain a data-driven set of results, our
22 primary analysis used a whole-brain approach.

23

24 **MATERIALS and METHODS**

1

2 *Participants*

3 A total of 23 college students participated in the study. One participant was
4 excluded from the analysis due to excessive head motion, and the remaining 22
5 participants were included in the analyses (11 female; mean age = 19.7, SD = 1.35).
6 All participants were pre-screened so that all of them knew at least 20 members of the
7 idol group and had at least one highly liked member and one highly disliked member.
8 The participants were all right-handed with no history of neurological or psychiatric
9 illness. All participants gave written informed consent for participation, and the study
10 was approved by the Institutional Review Board of Tamagawa University.

11

12 *Stimuli*

13 Photographs of the faces of individual members of the Japanese female idol group
14 “AKB48” (<https://en.wikipedia.org/wiki/AKB48>) were used in the present study. All
15 stimuli were obtained from the Internet. AKB48 consists of more than 100 members
16 and is popular especially among young Japanese people. We decided to use this idol
17 group for our source of experimental stimuli for two reasons: first, and most
18 importantly, their unique annual election contest allows us to have an incentive-
19 compatible choice task with real-world validity (see below for more detail on the
20 choice task). Every year, the idol group has a unique annual election contest as a
21 marketing strategy. Before the election contest, their new album is released, which
22 includes a voting code. Each fan can vote for his/her favorite member using the code,
23 and this election contest has a significant influence on each member’s media
24 exposure. The more votes a member receives, the more heavily she will be promoted.

1 This election system allows each fan to vote more than once by purchasing many
2 albums. Second, related to the first point, because of the competitive relationship
3 among members of the group created by the election contest, each fan (participant)
4 typically has a positive attitude toward some members and a negative attitude toward
5 other members. This large variance in attitudes provides an ideal real-world
6 background for the purpose of the current study.

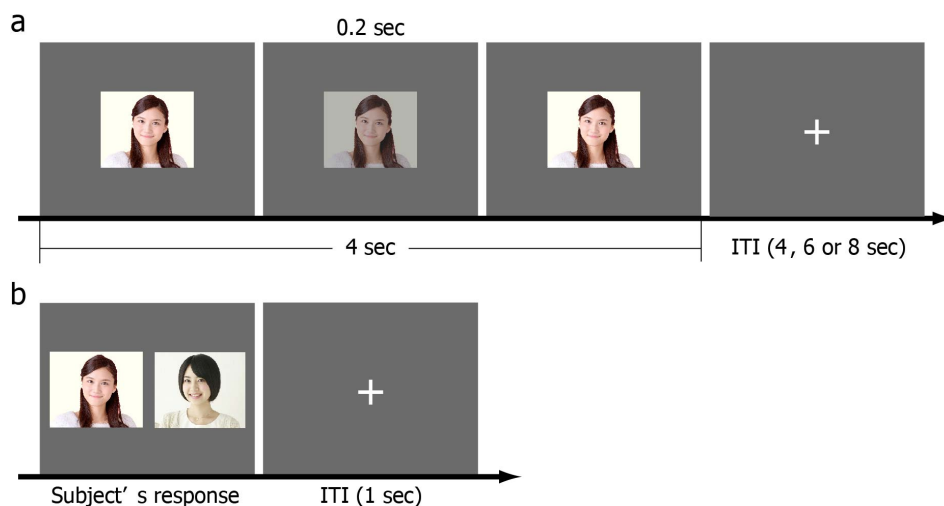
7 During the fMRI scanning, each participant was presented with pictures of 10
8 different members. Since we were interested in the neural representation of already
9 established social attitudes (as opposed to first impressions toward people never seen
10 before), we ensured that each participant was familiar with 10 members presented
11 during the fMRI scanning with the following procedure. Before an fMRI experiment,
12 each participant was asked to provide the experimenter with names of his/her most
13 and least favorite members (at least one for each) by email. Based on the names each
14 participant gave, the experimenter picked 10 members for the participant including
15 his/her favorite and least favorite members. If participants named fewer than 10
16 members, the experimenter selected other members based on general popularity. On
17 the day of the fMRI experiment, before they entered an fMRI scanner, each
18 participant was shown pictures of the 10 members and asked if they could identify all
19 of them. If there were any members that could not be identified, participants were
20 asked to pick other members they knew from a list of all members' names and
21 pictures. Accordingly, picture stimuli used in the fMRI experiment were different for
22 each participant, although there was overlap. More specifically, a total of 43 different
23 members of the idol group were used in the experiment, and each of them was
24 presented to at least one participant. 15 out of the 43 members (34.9%) were
25 presented to only one participant. The most consistently used member was presented

1 to 15 different participants. However, these 15 participants had idiosyncratic attitudes
 2 toward the member ranging from revealed attitude scores of 3 to 17 (possible range =
 3 0 - 18; see below for more information on the revealed attitude score). Thus, the
 4 results of the across-participant MVPA reported below are highly unlikely to be
 5 explained by decoding of person identity or face features (e.g., hair length).

6

7 *Experimental tasks*

8 In each trial, a picture of a member of the idol group was presented for 4 seconds,
 9 and participants were asked to perform a simple button press task during the fMRI
 10 session. At a random point between 1.5 to 3 seconds after the onset of the picture
 11 presentation, the picture became darker for 0.2 sec (Fig.1a). Participants were asked
 12 to press the button as soon as they detected the luminance change. The inter-trial
 13 interval (ITI) was set to 4, 6 or 8 seconds (pseudo-randomly determined). In each
 14 fMRI run, each of 10 members was presented 3 times (30 trials). Each run lasted 5
 15 min, and there were 8 fMRI runs in total. Importantly, before the scanning,
 16 participants were not told that the experiment was about attitudes, and they were not
 17 explicitly asked to think about their attitude toward each member during the scanning.



18

Figure 1. Experimental tasks. **(a)** A single trial during the fMRI scanning. Each participant was asked to press a key when the luminance of a picture changed. **(b)** A single trial of the choice task. After the fMRI session, each participant performed the choice task. In each trial, two members of the idol group were presented on the screen, and participants were asked to choose the member they want to vote for at the next election event. Note that due to copyright restrictions, the two individuals depicted in the pictures in this figure are not actual members of the idol group.

After the fMRI session, participants were asked to perform a choice task outside of the scanner (they were not told before the scanning that there would be the choice task). In each trial, two members from the 10 members used in the fMRI task were presented on the computer screen (Fig.1b), and participants were asked to select the one they want to vote for at the next election event by pressing one of two keys on the keyboard. There were 45 unique choice pairs, and each pair was presented twice (with switched positions) so that participants made 90 binary choices in total. During the choice task, each of the 10 members was presented 18 times. Therefore, each member could be selected a minimum of 0 times and maximum of 18 times. Since any attitude measure is considered to be good so long as it can predict relevant behaviors, this “revealed attitude score” for each member was used as labels in the subsequent MVPA analysis. Thus, in this project, we aim to predict individual's choice behavior (i.e., revealed attitude score) based on brain activations that would encode attitude, but without relying on any psychological measures of attitude (e.g., self-report).

Importantly, before they started the choice task, all participants were instructed that after the choice task, one trial would be selected randomly, and we would actually vote for the member the participant selected in that trial. All participants were shown a new music album including a voting code, and using the code, the experimenter actually voted for a member selected by the participant in a randomly-selected trial.

Finally, participants were asked to rate each of 10 members on attitude (how much do you like this member?) and attractiveness (how attractive do you think this member is) using a 9-point scale. At the end of the experiment, all participants received the album in addition to the monetary compensation (8,000 Japanese yen) for their participation.

fMRI data acquisition

All fMRI data were acquired using a Siemens 3.0 Tesla Trio scanner with a 32 channel phased array headcoil. For functional imaging, interleaved T2*- weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 34 contiguous 3.5-mm-thick trans-axial slices covering nearly the entire cerebrum (repetition time [TR] = 2,000 ms; echo time [TE] = 25 ms; flip angle [FA] = 90°; field of view [FOV] = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.5 mm). A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

fMRI data preprocessing

The fMRI data were analyzed using SPM8 (Wellcome Department of Imaging Neuroscience) implemented in MATLAB (MathWorks). Before data processing and statistical analysis, we discarded the first four volumes to allow for T1 equilibration. After correcting for differences in slice timing within each image volume, head motion was corrected. Following motion correction, the volumes were normalized to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each individual participant to the EPI template (resliced to a voxel size

1 of $3.0 \times 3.0 \times 3.5$ mm). These normalized data were used for the MVPA data
 2 analyses. For the univariate analysis, the normalized fMRI data were spatially
 3 smoothed with an isotropic Gaussian kernel of 4 mm (full-width at half-maximum).

4

5 *fMRI data analysis: searchlight MVPA*

6 In the across-participant MVPA analysis, we attempt to predict attitudes toward
 7 each of 10 idol group members based on the data obtained from all other participants.
 8 We first ran a conventional general linear model (GLM) analysis. In the GLM, each of
 9 10 members was separately modeled (duration = 4 sec). Button presses (duration = 0
 10 sec) and head motions were also included in the model as nuisance regressors.
 11 Contrast images for each member were created by using the data from all of the 8
 12 fMRI runs. These contrast images were used as input, and revealed attitude scores
 13 (how many times each member was chosen during the choice task; 0-18) were used as
 14 labels in the MVPA analysis.

15 MVPA was performed by using custom-made MATLAB in combination with
 16 LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). In order to predict the
 17 parametric variable of revealed attitude scores, we employed support vector
 18 regression (SVR; Drucker et al., 1997), as implemented in LIBSVM, with a linear
 19 kernel and a cost parameter of $c = 0.01$. This cost parameter was selected a priori
 20 following a previous study investigating value and salience signals in the brain (Kahnt
 21 et al., 2014). The regression MVPA was performed using a whole brain searchlight
 22 procedure (Kriegeskorte et al., 2006) with a radius of 3 voxels (maximum of 123
 23 voxels, and less at the boundaries of the brain). In each searchlight, accuracy at
 24 predicting revealed attitude scores for the 10 members was computed using leave-

one-participant-out cross validation. In each cross-validation, one participant was left out, and the SVR was performed using the data from all other participants and then tested on the left-out participant. This procedure was repeated for each participant (a total of 22 times; to test the robustness of our findings, we also ran a 4-fold cross-validation which replicated the main findings [see Supplementary Results and Table S2]). In each searchlight analysis, Spearman's rank-order correlations were computed between a participant's revealed attitude scores and their predicted attitudes; this correlation value was then assigned to the center voxel of the searchlight, resulting in an anatomical correlation map for each participant. The correlation values were Fisher z-transformed, spatially smoothed with an isotropic Gaussian kernel of 4 mm (full-width at half-maximum), and then submitted to the second level analysis (i.e., one sample t-tests across all participants).

In order to identify the neural representations of attitude extremity, we also ran the same across-participant MVPA analysis using the attitude extremity score, which is computed by calculating the absolute value of the difference between a revealed attitude score for each member and the midpoint of the revealed attitude score (i.e., 9). We further ran three control MVPA analyses using 1) self-report attractiveness ratings, 2) self-report attitude ratings, and 3) reaction times (RTs) (see Supplementary Information).

fMRI data analysis: Univariate analysis

We also ran a standard univariate fMRI analysis to see whether univariate activations might be correlated with revealed attitude score. The GLM included three main regressors; 1) all idol group member presentations (duration = 4 sec), 2) member

1 presentations parametrically modulated by participant's revealed attitude score, and 3)
 2 member presentations parametrically modulated by participant's attractiveness rating.
 3 Button presses (duration = 0 sec) and head motions were also included in the model
 4 as nuisance regressors.

5 For both the MVPA and univariate analyses, the statistical threshold was set at $p <$
 6 0.001 voxelwise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple
 7 comparisons).

8

9 **RESULTS**

10 *Behavioral results*

11 Not surprisingly, revealed attitude scores (choice behavior) were highly
 12 correlated with self-reported attitudes (average $r = 0.91$, $t(21) = 18.11$, $p < 0.001$).
 13 They were also correlated with attractiveness ratings (average $r = 0.76$, $t(21) = 10.80$,
 14 $p < 0.001$). The revealed attitude scores were more strongly correlated with the
 15 attitude ratings than the attractiveness ratings ($t(21) = 3.33$, $p = 0.003$). Attitude
 16 ratings and attractiveness ratings were also correlated with each other (average $r =$
 17 0.76 , $t(21) = 10.69$, $p < 0.001$).

18 Inside the fMRI scanner, participants were asked to press the button as soon as a
 19 picture gets dimmed, and their performance for this simple button press task was
 20 nearly perfect (98.6%), and average RT across participants was 297 ms (SD = 52),
 21 indicating that participants paid attention to each picture stimulus. The analyses also
 22 revealed that RTs were significantly negatively correlated with participants' revealed
 23 attitude scores (average $r = -0.20$, $t(21) = -3.14$, $p = 0.005$) and attitude ratings
 24 (average $r = -0.16$, $t(21) = 2.27$, $p = 0.034$). Thus, the more favorable their attitudes

1 were toward members, the faster (smaller) the RT. This result may suggest that
 2 pictures of their favorite members captured attention relative to less favorite
 3 members, which in turn enhanced their reaction times. RTs were not correlated with
 4 the attractiveness ratings (average $r = -0.08$, $t(21) = 1.24$, $p = 0.23$, n.s.) or the attitude
 5 extremity scores (average $r = -0.12$, $t(21) = 1.30$, $p = 0.21$, n.s.).

6 After the scanning, participants were asked to make binary choices between two
 7 members of the idol group for whom they wanted to vote in the next election event.
 8 They were presented with the same choice pair twice, and their choices were largely
 9 consistent across two presentations of the same pair (choice consistency = 86.7%).
 10 Inconsistent choices were more likely to happen when the difference in attitude
 11 ratings between two members was small (e.g., when two persons were similarly
 12 liked). When the within-pair ratings difference was equal to or less than 2, choice
 13 consistency was 79.3%, while it was 94.0% when the difference was more than 2.
 14 This difference between the two choice consistency values was significant ($t(21) =$
 15 5.32 , $p < 0.001$). Not surprisingly, participants' choices were highly accurately
 16 predicted by their self-report attitudes (mean choice prediction accuracy = 88.5%),
 17 and the accuracy was significantly higher than chance (50%) ($t(21) = 20.1$, $p < 0.001$;
 18 note that choice accuracy was computed after excluding inconsistent choice pairs).

19

20 *fMRI results: searchlight MVPA*

21 The searchlight MVPA analysis revealed that spatial activation patterns in the
 22 anterior part of the right striatum significantly predicted participants' revealed attitude
 23 scores (Fig. 2a). The average Spearman's correlation coefficient in the peak of this
 24 anterior striatum cluster was 0.26 ($t(21) = 7.24$, $p < 0.001$). Activation patterns in

1 right inferior frontal gyrus (IFG) also significantly predicted participants' attitudes
 2 (Spearman's $Rho = 0.23$, $t(21) = 6.49$, $p < 0.001$; Fig. 2a, Table 1). No other region
 3 significantly predicted participants' revealed attitude scores. The average correlation
 4 of 0.26 found in the anterior striatum is equivalent to 59.6% in terms of accuracy for
 5 predicting participants' binary choices (see Supplementary Results for more details
 6 about the choice prediction accuracy results and the direct comparison of the choice
 7 prediction accuracy between neural and self-report measures).

8 The MVPA analysis with the attitude extremity score revealed that activation
 9 patterns in a more posterior part of the right striatum significantly predicted attitude
 10 extremity (Fig. 2a). Activation patterns in other brain regions including inferior
 11 orbitofrontal cortex (OFC), supplementary motor area (SMA), posterior insula,
 12 precentral gyrus, inferior parietal lobule and occipital pole also significantly predicted
 13 attitude extremity (Table 1).

14 To further quantify the functional dissociation within the striatum, we extracted
 15 each participant's correlation coefficients for both the revealed attitude score and the
 16 attitude extremity score from the voxels within a 4mm sphere surrounding the peaks
 17 of these two striatum sub-regions (note that the following statistical analysis is of
 18 course not statistically independent of our above discovery; it is intended merely to
 19 provide further detail for the dissociation). After Fisher's z transforming these
 20 correlation coefficients, we performed a 2 (anterior vs. posterior striatum) \times 2
 21 (revealed attitude score vs. attitude extremity score) repeated-measure analysis-of-
 22 variance (ANOVA). It revealed a significant interaction ($F(1, 21) = 26.4$, $p < 0.001$)
 23 (Fig. 2b). We further performed a paired t -test within each of the anterior and
 24 posterior striatum regions. In the anterior striatum, prediction performance (mean
 25 correlation) was significantly higher for revealed attitude compared to attitude

1 extremity ($t(21) = 4.24, p < 0.001$). In contrast, in the posterior striatum, prediction
2 performance was significantly higher for attitude extremity than for revealed attitude
3 ($t(21) = 3.48, p = 0.002$) (Fig. 2b). Furthermore, correlations were not significantly
4 different from zero for attitude extremity in the anterior striatum and for revealed
5 attitude in the posterior striatum. These results indicate a clear functional dissociation
6 within the striatum such that the anterior striatum represents an individual's
7 evaluation of each celebrity, whereas the posterior striatum represents how extreme
8 his/her attitude is regardless of valence. Main effects of region and attitude property
9 were both not significant ($ps > 0.50$). Finally, a mixed ANOVA with gender as an
10 additional between-subject factor revealed no main or interaction effect involving
11 gender (all $ps > 0.35$).

12 Our control analyses further confirmed that the MVPA results reported above
13 cannot be explained by perceived attractiveness of faces or reaction times (i.e.,
14 attention) (see Supplemental Results, Figure S1 and Table S1). Furthermore,
15 additional searchlight MVPA analysis restricted to an anatomical mask of the striatum
16 confirmed that activation patterns within the striatum are responsible for the findings
17 reported above, and could not be ascribed to partial volume effects from nearby
18 regions (see Supplemental Results).

19

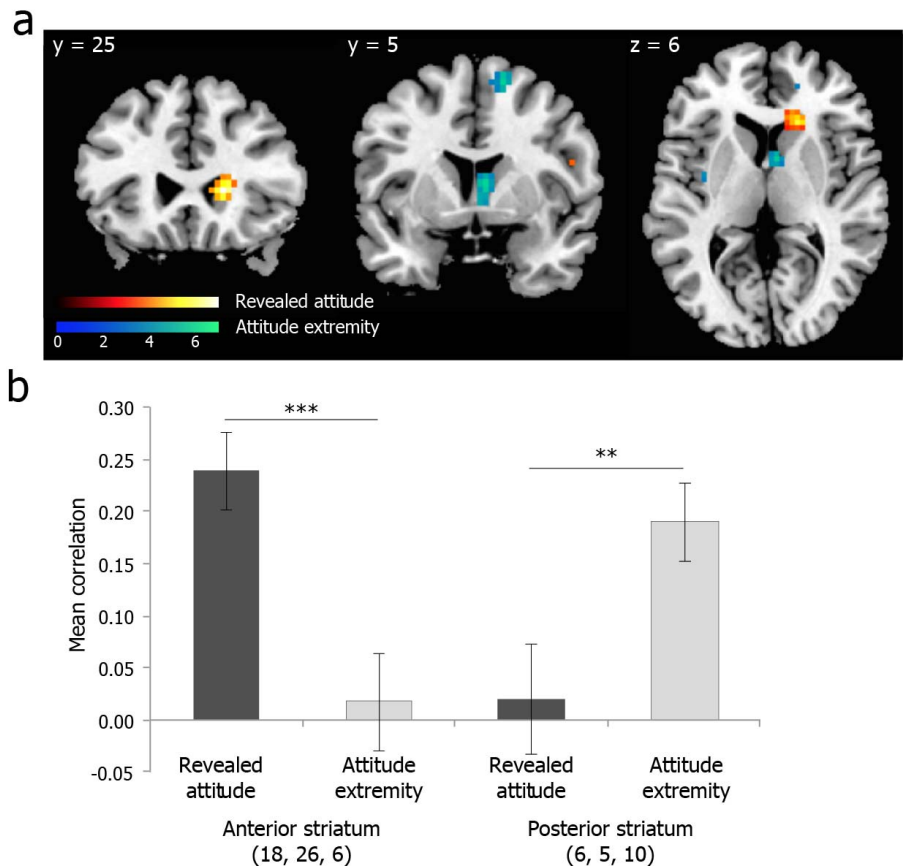


Figure 2: fMRI results. (a) Activation patterns in the anterior striatum significantly predicted participants' revealed attitude scores, whereas activation patterns in the posterior striatum (blue) predicted attitude extremity. **(b)** Mean within-participant Spearman's rank correlation between predicted values and actual values (revealed attitude or attitude extremity) from the anterior and posterior striatum. ** $p < 0.01$, *** $p < 0.001$. Error bars represent standard error of mean (SEM).

Table 1. Regions where activity patterns significantly predicted revealed attitude score and attitude extremity score.

Location	Side	MNI coordinates			z value	Number of Voxel
		x	y	z		
<i>Revealed attitude score</i>						
Anterior striatum	R	18	26	6	5.07	37
Inferior frontal gyrus (IFG)	R	54	14	20	4.76	56
<i>Attitude extremity score</i>						
posterior striatum	R	6	5	10	4.56	31
inferior orbitofrontal cortex (OFC)	R	30	38	-4	3.90	25
supplementary motor area (SMA)	R	15	2	66	4.82	26
posterior insula	L	-48	-7	2	4.00	46
precentral gyrus	L	-33	-19	62	5.07	63
inferior parietal lobule	L	-51	-31	44	4.38	48
occipital pole	R	30	-95	-15	4.48	45

Note that the size of a voxel is $3 \times 3 \times 3.5$ mm

fMRI results: Univariate analysis

The results of the univariate fMRI data analysis revealed that the activity only in the left posterior fusiform gyrus ($x = -27$, $y = -91$, $z = -15$; 60 voxels) was significantly positively correlated with the revealed attitude scores. Only when the threshold was lowered to $p < 0.005$, did we find activations in the ventromedial prefrontal cortex (vmPFC; $x = -6$, $y = 53$, $z = -12$), one of the areas commonly associated with preference and valuation (see Bartra et al., 2013; Clithero and Rangel, 2014). We did not find any activation in the striatum even with this lowered threshold. In addition, no area was significantly negatively correlated with the revealed attitude scores. Furthermore, no area was significantly correlated (either positively or negatively) with the attitude extremity scores.

1

2 **DISCUSSION**

3 The present study investigated the possibility of measuring people's social attitude
4 toward familiar others based on multivariate neural activations from fMRI. The
5 across-participant MVPA revealed that activation patterns in the anterior striatum can
6 significantly predict the choices made based on one's attitude toward members of an
7 idol group. This result indicates that spatial patterns of activations in the anterior
8 striatum contain reliable information about an individual's attitudes, and these neural
9 representations of attitudes in the anterior striatum are sufficiently similar across
10 different individuals so that it is possible to infer attitudes of an individual based on
11 the association between revealed attitudes and brain activation patterns found in other
12 individuals. In contrast, our univariate analysis failed to find any significant
13 activations related to participants' attitudes (and attitude extremity) in reward-related
14 areas, indicating that average amplitude of activity is an insufficiently sensitive
15 measure to represent attitudes in the present study. Although previous studies have
16 demonstrated that univariate activations in reward-related brain regions such as
17 vmPFC and striatum are correlated with people's attitudes or preference for various
18 items, especially when individuals were asked to report their preference for each
19 stimulus inside the scanner (Izuma et al., 2010; Lebreton et al., 2009; for review, see
20 Bartra et al., 2013), such univariate (mean) activation does not appear to robustly
21 track preference for stimuli during passive viewing (but see also Levy et al., 2011 and
22 Tusche et al., 2013). In general, our results add to a growing body of evidence
23 showing the higher sensitivity of MVPA compared with univariate analyses (e.g.,
24 Jimura & Poldrack, 2012; Kohler et al., 2013).

Although we found that activation patterns in the anterior striatum and right IFG can predict participants' attitudes revealed in the choice task (i.e., behavior) significantly better than would be expected by chance, prediction accuracy was considerably lower than their self-reported attitude ratings. Thus, at this point, self-report measure outperforms neural measures of attitudes. This result is not surprising because self-report attitude toward celebrities should be much less susceptible to social desirability bias compared to attitudes toward socially sensitive issues (e.g., racial prejudice), and there is no apparent reason for participants to regulate their answers during attitude ratings of the celebrities. Nonetheless, our present study provides an important reference point to which future social neuroscience studies can be compared. For example, it might well be the case that neurally measured attitude can outperform self-report or implicit measures of attitudes in predicting racially discriminatory behavior for which prediction accuracy of behavioral measures (both implicit and explicit measures) are known to be limited (Greenwald et al., 2009; Oswald et al., 2013), an important direction for future studies.

The greatest potential of a neural measure of attitude comes from the fact that we could successfully predict participants' choice behavior (i.e., revealed attitude score) from their brain activations alone without asking them to engage in any attitude-related task during scanning, and without incorporating information about attitude judgments into our analysis of the fMRI data. While implicit attitude measures circumvent some of the problems with self-reports, they cannot be "process-pure", so that a score on an implicit attitude measure generally reflects some factors other than implicit attitude toward a stimulus (Conrey et al., 2005; Sherman, 2009). In contrast, in the present study, participants were never asked to report their attitudes (i.e., self-report) or make any response based on attributes of attitude objects (i.e., implicit

1 measure) during the scanning – yet the activation patterns could robustly predict later
2 choices based on attitudes. This suggests that it may be possible to infer people’s
3 spontaneous attitudes using incidental brain imaging methods. As neurally measured
4 incidental attitudes are unlikely to be influenced by any automatic or controlled
5 processes, there is no translational gap between the construct (i.e., attitude) and the
6 way it is measured (Sherman, 2009). Our findings suggest that it may be possible to
7 measure people’s attitudes even when they are unable and/or unwilling to report them
8 truthfully (although ethical concerns for such an approach need to be carefully
9 considered). Thus, a neural measure of incidental attitudes has the potential to be
10 highly, if not completely, process-pure as well as relatively effort-free, offering some
11 distinct advantages over any implicit (and explicit) behavioral measures.

12 It should be noted that while the neural measure was outperformed by self-report
13 in predicting individual’s choice behaviors, the choice prediction accuracy found in
14 the present study (59.6%) is comparable to the three past neuroeconomics studies
15 (Tusche et al., 2010; Levy et al., 2011; Smith et al., 2014), which attempted to predict
16 individual’s choices from neural responses while passively viewing stimuli (Levy et
17 al., 2011; Smith et al., 2014) or actively engaging an demanding attention task
18 (Tusche et al., 2010). In contrast to the present study, all of the three studies focused
19 on within-participant predictions (i.e., predicting a participant’s choice between two
20 items based on his/her brain activations in response to other items) and reported the
21 choice prediction accuracy ranging from 56% to 83%. Smith et al. (2014) also tested
22 across-group predictions, which is conceptually similar to our across-participant
23 predictions and reported the prediction accuracy of 61.2%. Taken together with the
24 findings of Smith et al. (2014), the present results indicate that neural activation
25 patterns associated with attitudes are sufficiently similar across different individuals,

1 suggesting a great potential for an objective measure of attitudes using a
2 neuroimaging method.

3 We also found that attitude extremity can be predicted by activation patterns in
4 more posterior parts of the striatum among other regions. This result indicates that
5 both highly liked and highly disliked members induce similar patterns of activations
6 in this region. While the striatum is widely implicated in reward processing (Delgado,
7 2007), it is also known to respond to saliency (e.g., Zink et al., 2004; Jensen et al.,
8 2007). As stimuli with high attitude extremity scores in the present study are also
9 likely to be highly salient, our results are consistent with the role of the striatum in
10 processing saliency. In addition, the pattern of results we obtained (Fig. 2b) clearly
11 indicates a functional dissociation within the striatum; whereas the anterior striatum
12 represent attitudes, the posterior striatum represents saliency or attitude extremity.
13 Functional dissociation within the striatum found in the present study is further
14 supported by past studies which identified several sub-regions within the striatum
15 based on intrinsic functional connectivity (Choi et al., 2012) and patterns of
16 coactivation with other cortical areas (Pauli et al., 2016). Although the reports on how
17 the striatum is organized differ slightly across studies, it seems clear that the two
18 striatal regions found in the present study lie in different sub-regions within the
19 striatum. Furthermore, the anterior striatum's role in representing attitudes is
20 consistent with the findings from large-scale coactivation data that this sub-region is
21 the most strongly involved in representing stimulus value (Pauli et al., 2016) (note
22 that the anterior striatum cluster found in the present study seems to be the closest to
23 the cluster labeled “ventral striatum” in Pauli et al., 2016). In contrast, however,
24 posterior striatum identified in the present study seems to be the striatal sub-region
25 most strongly associated with executive function (Pauli et al., 2016), which is not

1 necessarily consistent with our finding that this region is involved in attitude
2 extremity or salience.

3 Nonetheless, our findings may be explained by two types of dopamine neurons
4 systematically located in the midbrain and projections from the midbrain to the
5 striatum. Matsumoto and Hikosaka (2009) found two types of dopamine neurons in
6 the monkey midbrain; those that are excited by positive stimuli (juice) and inhibited
7 by negative stimuli (airpuff to the eyes), and those that are excited by both the
8 positive and negative stimuli (Matsumoto and Hikosaka, 2009). Consistently across
9 two monkeys, those neurons which respond to both positive and negative stimuli (i.e.,
10 motivationally salient stimuli) are located in the dorsolateral substantia nigra pars
11 compacta which projects mainly to the dorsal striatum, whereas those neurons which
12 predominantly respond to positive stimuli are located in the ventromedial substantia
13 nigra pars compacta (also ventral tegmental area) which send projections mainly to
14 the ventral striatum (Haber and Knutson, 2010). Thus, striatal functional dissociation
15 found in the present study might reflect the activations of these two types of neurons
16 in the midbrain. Thus, although speculative, the across-participant MVPA results may
17 suggest that different populations of neurons encode reward (attitude) and saliency
18 (attitude extremity), each of which is localized in a different sub-region of the
19 striatum and receives projection from dopamine neurons in a different sub-region of
20 the midbrain.

21

22 **CONCLUSION**

23 The present study investigated the potential of a neuroimaging method to predict
24 people's incidental social attitudes toward others. We found that patterns of

1 activations in the anterior striatum can reliably predict an individual's attitudes toward
2 famous people, suggesting the feasibility of such an approach. Although the
3 prediction accuracy was higher for the self-report measure than the neural measure,
4 the present study represents an essential first step toward neural measures of social
5 attitudes and demonstrated that we could successfully predict attitudes without asking
6 participants to engage in any attitude-related task. While we focused on explicit
7 attitudes toward familiar people, it will be important in future research to see how
8 accurately similar neural measures might predict implicit attitudes such as prejudice
9 toward racial or social outgroups.

10

REFERENCES

- 1
- 2 Amodio, D. M. (2014) The neuroscience of prejudice and stereotyping. *Nature*
- 3 *Reviews Neuroscience*, 15, 670-682.
- 4 Arkes, H. R., & Tetlock, P. E. (2004) Attributions of implicit prejudice, or "would
- 5 Jesse Jackson 'fail' the implicit association test?". *Psychological Inquiry*, 15,
- 6 257-278.
- 7 Bartra, O., McGuire, J. T., & Kable, J. W. (2013) The valuation system: a coordinate-
- 8 based meta-analysis of BOLD fMRI experiments examining neural correlates
- 9 of subjective value. *NeuroImage*, 76, 412-427.
- 10 Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006) Decoding the implicit
- 11 association test: Implications for criterion prediction. *Journal of Experimental*
- 12 *Social Psychology*, 42, 192-212.
- 13 Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000) Stalking the perfect
- 14 measure of implicit self-esteem: The blind men and the elephant revisited?
- 15 *Journal of Personality and Social Psychology*, 79, 631-643.
- 16 Buchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998) Brain systems mediating
- 17 aversive conditioning: an event-related fMRI study. *Neuron*, 20, 947-957.
- 18 Choi, E. Y., Yeo, B. T., & Buckner, R. L. (2012) The organization of the human
- 19 striatum estimated by intrinsic functional connectivity. *Journal of*
- 20 *Neurophysiology*, 108, 2242-2263.
- 21 Clithero, J. A., & Rangel, A. (2014) Informatic parcellation of the network involved
- 22 in the computation of subjective value. *Social Cognitive and Affective*
- 23 *Neuroscience*, 9, 1289-1302.
- 24 Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005)
- 25 Separating multiple processes in implicit social cognition: The quad model of

- 1 implicit task performance. *Journal of Personality and Social Psychology*, 89,
2 469-487.
- 3 Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008) Affective flexibility:
4 Evaluative processing goals shape amygdala activity. *Psychological Science*,
5 19, 152-160.
- 6 Cunningham, W. A., Johnson, M. K., Gatenby, J. C., Gore, J. C., & Banaji, M. R.
7 (2003) Neural components of social evaluation. *Journal of Personality and*
8 *Social Psychology*, 85, 639-649.
- 9 Delgado, M. R. (2007) Reward-related responses in the human striatum. *Annals of the*
10 *New York Academy of Sciences*, 1104, 70-88.
- 11 Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000) Tracking
12 the hemodynamic responses to reward and punishment in the striatum. *Journal*
13 *of Neurophysiology*, 84, 3072-3077.
- 14 DeMaio, T. J. (1984) Social desirability and survey measurement: A review. In
15 *Surveying subjective phenomena* (Turner CF, Martin E, eds). New York:
16 Russell Sage.
- 17 Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997) Support
18 vector regression machines. *Advances in Neural Information Processing*
19 *Systems*, 9, 155-161.
- 20 Etkin, A., Egner, T., & Kalisch, R. (2011) Emotional processing in anterior cingulate
21 and medial prefrontal cortex. *Trends in Cognitive Sciences*, 15, 85-93.
- 22 Fazio, R. H., & Olson, M. A. (2003) Implicit measures in social cognition research:
23 Their meaning and use. *Annual Review of Psychology*, 54, 297-327.
- 24 Fiedler, K., Messner, C., & Bluemke, M. (2006) Unresolved problems with the “I”,
25 the “A”, and the “T”: A logical and psychometric critique of the Implicit

- 1 Association Test (IAT). *European Review of Social Psychology*, 17, 74-147.
- 2 Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998) Measuring individual
3 differences in implicit cognition: The implicit association test. *Journal of*
4 *Personality and Social Psychology*, 74, 1464-1480.
- 5 Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji M. R. (2009)
6 Understanding and Using the Implicit Association Test: III. Meta-Analysis of
7 Predictive Validity. *Journal of Personality and Social Psychology*, 97, 17-41.
- 8 Haber, S.N., & Knutson, B. (2010) The reward circuit: Linking primate anatomy and
9 human imaging. *Neuropsychopharmacology*, 35, 4-26.
- 10 Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008)
11 Dissociating the role of the orbitofrontal cortex and the striatum in the
12 computation of goal values and prediction errors. *Journal of Neuroscience*, 28,
13 5623-5630.
- 14 Ito, T. A., & Bartholow, B. D. (2009) The neural correlates of race. *Trends in*
15 *Cognitive Sciences*, 13, 524-531.
- 16 Izuma, K., Saito, D. N., & Sadato, N. (2008) Processing of social and monetary
17 rewards in the human striatum. *Neuron* 58, 284-294.
- 18 Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto,
19 K. (2010) Neural correlates of cognitive dissonance and choice-induced
20 preference change. *Proceedings of National Academy of Science USA*, 107,
21 22014-22019.
- 22 Jensen, J., Smith, A. J., Willeit, M., Crawley, A. P, Mikulis, D. J., Vitcu, I., & Kapur,
23 S. (2007) Separate brain regions code for salience vs. valence during reward
24 prediction in humans. *Human Brain Mapping*, 28, 294-302.
- 25 Jimura, K., & Poldrack, R. A. (2012) Analyses of regional-average activation and

- 1 multivoxel pattern information tell complementary stories. *Neuropsychologia*
- 2 50, 544-552.
- 3 Kahnt, T., Park, S. Q., Haynes, J. D., & Tobler, P. N. (2014) Disentangling neural
- 4 representations of value and salience in the human brain. *Proceedings of*
- 5 National Academy of Science USA, 111, 5000-5005.
- 6 Karpinski, A., & Hilton, J. L. (2001) Attitudes and the Implicit Association Test.
- 7 *Journal of Personality and Social Psychology*, 81, 774-788.
- 8 Karpinski, A., Steinman, R. B., & Hilton J. L. (2005) Attitude importance as a
- 9 moderator of the relationship between implicit and explicit attitude measures.
- 10 *Personality and Social Psychology Bulletin*, 31, 949-962.
- 11 Knight, D. C., Nguyen, H. T., & Bandettini, P. A. (2005) The role of the human
- 12 amygdala in the production of conditioned fear responses. *NeuroImage*, 26,
- 13 1193-1200.
- 14 Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000) FMRI visualization of
- 15 brain activity during a monetary incentive delay task. *NeuroImage*, 12, 20-27.
- 16 Knutson KM, Wood JN, Spampinato MV, Grafman J (2006) Politics on the brain: an
- 17 FMRI investigation. *Social Neuroscience*, 1, 25-40.
- 18 Kohler, P. J., Fogelson, S. V., Reavis, E. A., Meng, M., Guntupalli, J. S., Hanke, M.,
- 19 Halchenko, Y. O., Connolly, A. C., Haxby, J. V., & Tse, P. U. (2013) Pattern
- 20 classification precedes region-average hemodynamic response in early visual
- 21 cortex. *NeuroImage*, 78, 249-260.
- 22 Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006) Information-based functional
- 23 brain mapping. *Proceedings of National Academy of Science USA*, 103, 3863-
- 24 3868.
- 25 Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012) The neuroscience of race. *Nature*

- 1 Neuroscience, 15, 940-948.
- 2 LaBar, K. S., Gatenby, J. C., Gore, J.C., LeDoux, J. E., & Phelps, E. A. (1998)
- 3 Human amygdala activation during conditioned fear acquisition and
- 4 extinction: A mixed-trial fMRI study. *Neuron*, 20, 937-945.
- 5 Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009) An
- 6 automatic valuation system in the human brain: Evidence from functional
- 7 neuroimaging. *Neuron*, 64, 431-439.
- 8 Levy, D. J., & Glimcher, P. W. (2012) The root of all value: A neural common
- 9 currency for choice. *Current Opinion in Neurobiology*, 22, 1027-1038.
- 10 Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011) Choice from non-
- 11 choice: Predicting consumer preferences from blood oxygenation level-
- 12 dependent signals obtained during passive viewing. *Journal of Neuroscience*,
- 13 31, 118-125.
- 14 Luttrell, A., Stillman, P. E., Hasinski, A. E., & Cunningham, W. A. (2016) Neural
- 15 dissociations in attitude strength: Distinct regions of cingulate cortex track
- 16 ambivalence and certainty. *Journal of Experimental Psychology: General*, 145,
- 17 419-433.
- 18 Matsumoto, M., & Hikosaka, O. (2009) Two types of dopamine neuron distinctly
- 19 convey positive and negative motivational signals. *Nature*, 459, 837-U834.
- 20 O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001)
- 21 Abstract reward and punishment representations in the human orbitofrontal
- 22 cortex. *Nature Neuroscience*, 4, 95-102.
- 23 Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013)
- 24 Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion
- 25 studies. *Journal of Personality and Social Psychology*, 105, 171-192.

- 1 Pauli, W. M., O'Reilly, R. C., Yarkoni, T., & Wager, T. D. (2016) Regional
2 specialization within the human striatum for diverse psychological functions.
3 Proceedings of National Academy of Science USA, 113, 1907-1912.
- 4 Petty, R. E., & Krosnick, J. A. editors. (1995). Attitude strength: Antecedents and
5 consequences, Mahwah, NJ: Erlbaum.
- 6 Petty R. E., Fazio R. H., & Brinol P., editors (2009) Attitudes: Insights from the new
7 implicit measures. New York: Psychology Press.
- 8 Phelps E. A., O'Connor K. J., Cunningham W. A., Funayama E. S., Gatenby J. C.,
9 Gore J. C., Banaji M. R. (2000) Performance on indirect measures of race
10 evaluation predicts amygdala activation. Journal of Cognitive Neuroscience,
11 12, 729-738.
- 12 Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003) Common
13 method biases in behavioral research: A critical review of the literature and
14 recommended remedies. Journal of Applied Psychology, 88, 879-903.
- 15 Powell, M. C., & Fazio, R. H. (1984) Attitude accessibility as a function of repeated
16 attitudinal expression. Personality and Social Psychology Bulletin, 10, 139-
17 148.
- 18 Sescousse, G., Caldu, X., Segura, B., & Dreher, J. C. (2013) Processing of primary
19 and secondary rewards: A quantitative meta-analysis and review of human
20 functional neuroimaging studies. Neuroscience & Biobehavioral Reviews, 37,
21 681-696.
- 22 Sherman, J. W. (2009) Controlled influences on implicit measures: Confronting the
23 myth of process-purity and taming the cognitive monster. In Attitudes: Insights
24 from the new implicit measures (Petty RE, Fazio RH, Brinol P, eds). New
25 York, Hove: Psychology Press.

- 1 Smith, A., Bernheim, B. D., Camerer, C. F., Rangel, A. (2014) Neural Activity
2 Reveals Preferences without Choices. American Economic Journal:
3 Microeconomics, 6, 1-36.
- 4 Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011) Implicit race
5 attitudes predict trustworthiness judgments and economic trust decisions.
6 Proceedings of National Academy of Science USA, 108, 7710-7715.
- 7 Tusche, A., Bode, S., & Haynes, J. D. (2010) Neural responses to unattended products
8 predict later consumer choices. Journal of Neuroscience, 30, 8024-8031.
- 9 Tusche, A., Kahnt, T., Wisniewski, D., & Haynes, J. D. (2013) Automatic processing
10 of political preferences in the human brain. NeuroImage, 72, 174-182.
- 11 Van Boven, L., Judd, C. M., & Sherman, D. K. (2012) Political polarization
12 projection: Social projection of partisan attitude extremity and attitudinal
13 processes. Journal of Personality and Social Psychology, 103, 84-100.
- 14 Vickery, T. J., Chun, M. M., & Lee, D. (2011) Ubiquity and specificity of
15 reinforcement signals throughout the human brain. Neuron, 72, 166-177.
- 16 Wittenbrink, B., & Schwarz, N., editors. (2007) Implicit measures of attitudes:
17 Procedures and controversies. New York: Guilford.
- 18 Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., Chappelow, J. C., Berns, G. S.
19 (2004) Human striatal responses to monetary reward depend on saliency.
20 Neuron, 42, 509-517.